图书情报工作
第 65 卷 第 14 期 2021 年 7 月

基于典籍文本的农作物时间分布及演化特征研究*
——以《食货志》为例

■ 崔斌[1,2] 王东波[1,2] 黄水清[1,2]
[1] 南京农业大学信息管理学院 南京 210095 [2]南京农业大学人文与社会计算研究中心 南京 210095

摘 要：[目的/意义]我国农作物种植历史悠久，分析古代农作物的时间分布与发展演化情况对优化现代农业种植结构具有重要意义。[方法/过程]提出一套深入典籍文本内容的农作物时间分布及演化特征分析方法流程，主要包括语料获取与数字化、分词与实体关系抽取、时间分布特征分析、演化特征分析 4 部分，并选取 15 本史书中的《食货志》文本进行实证分析。[结果/结论]基于《食货志》文本的分析结果得到历史学、经济学、文献学等多学科相关研究资料的佐证，验证了方法的可行性与有效性，可以为基于典籍文本的古代农作物时间分布及演化特征分析提供借鉴。但未来还需要在提高自动化水平、扩大研究样本、细化事件类型等方面进一步优化方法流程。

关键词：实体关联 数字人文 食货志 农作物 可视化
分类号：G251
DOI：10.13266/j.issn.0252－3116.2021.14.011

1 引言

农为邦本，本固邦宁。农业是国民经济发展的重要基础，习近平在中央农村工作会议上强调："要坚持把解决好'三农'问题作为全党工作重中之重"[1]。我国有着悠久的农业种植历史，大约在一万年前就已经开始种植谷物[2]。从最初"凡可食之物皆可植"的"百谷"时期，到后来的"九谷""五谷"发展，再到玉米、番薯、马铃薯等域外作物的大量引进，我国农作物类型和种植技术不断完善，农业发展也逐步走向成熟。从历史的角度分析我国农作物的起源与发展，揭示农作物在不同历史时期的种植规律，对于现代农作物种质资源（又称遗传资源）收集与品种改良都具有重要意义。因此，研究人员从农学、历史学、考古学、文献学等多个视角，依据考古发现及史书、地方志等古代典籍文献，对农作物种植的历史演变情况展开研究。

李成[3]通过梳理史前到两汉时期黄河流域小麦相关资料，分析了小麦种植的时间分布情况。刘兴林[4]通过调研各地区考古文献的记录信息，分析了先秦时期粟、黍、稻、小麦等农作物的时间分布特征。简思敏等[5]基于区县地方志文献资料，分析了明清时期福建地区水稻、茶叶等作物的时间分布规律。李静[6]以北川地区的县志资料为基础，分析了清至民国时期该地区农作物的种植、传播和分布变化情况。上述研究主要以单一朝代或相距较近的几个朝代为研究区间，部分学者进一步在较长的历史时期内研究农作物的演化脉络。朱睿等[7]依据出土文物、古籍记载等分析了苎麻作物在中国的起源、分布和栽培利用历史。周跃中[8]以《汉书》《氾胜之书》等史书典籍为基础，归纳分析了从先秦到明清时期的农作物种类及其演变情况。彭景元[9]基于《史记》《汉书》和地方志等典籍资料，从农业起步、农业开发、海外作物引进等层面分析了我国闽南地区的农业发展历程。

现有研究主要以人工解读的方式从史料记载中推演农作物演变情况，分析结果较为准确。但这类方法主要存在 3 点不足：①一般要求分析人员同时具备一定的农学和历史学知识，受人员知识结构的影响较大；②往往只针对单一或简单几种农作物展开分析，较少同时对多类型农作物进行考量；③对史料的通读和分析耗时耗力，分析效率有待提升。而随着数据挖掘技

* 本文系国家社会科学基金重大项目"基于《汉学引得丛刊》的典籍知识库构建及人文计算研究"（项目编号：15ZDB127）研究成果之一。
作者简介：崔斌（ORCID：0000－0002－6877－7030），博士研究生；王东波（ORCID：0000－0002－9894－9550），教授，博士生导师；黄水清（ORCID：0000－0002－1646－9300），教授，博士生导师，通讯作者，E-mail：sqhuang@njau.edu.cn。
收稿日期：2020－12－10 修回日期：2021－04－17 本文起止页码：90－100 本文责任编辑：易飞

术的不断发展与科学计量方法的日益多样化，利用自然语言处理技术深入古籍文本内容进行知识挖掘就成为有效解决这些问题的重要手段，其分析结果可与人工解读结果相互佐证、相互补充。基于此，本文提出基于典籍文本的农作物时间分布及演化特征分析方法流程，一方面从时间层面上分析我国古代农作物在不同朝代和年号阶段中的整体分布情况（农作物时间分布），另一方面从演化层面上分析我国古代农作物分布随时间推移的演变规律，并以《食货志》为例对其可行性与有效性进行验证。

2 分析方法与流程

本文提出一套深入古代典籍文本内容的农作物时间分布及演化特征分析方法流程，主要包括语料获取与数字化、分词与实体关系抽取、时间分布特征分析、演化特征分析4部分内容，如图1所示：

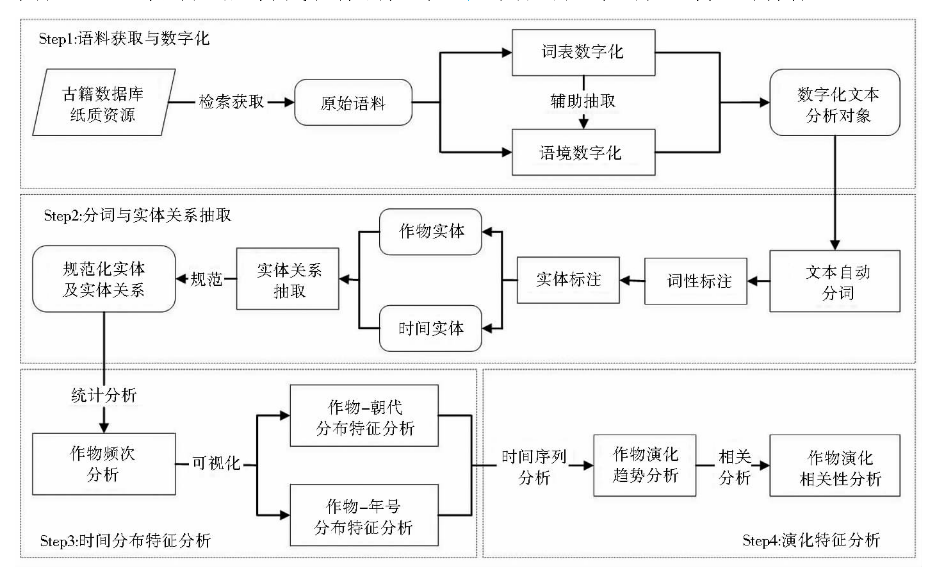


图1 农作物时间分布及演化特征分析方法流程

2.1 语料获取与数字化

语料获取与数字化是分析流程的第一步，也是后续各种分析的重要基础。原始语料一般包括词表语料与原文语料两类：词表语料以引得类文献为主，其作用是从中抽取词汇底表作为构建领域词表的重要基础，该类文献多为纸质资源；原文语料即古籍原文全文资料，常见的收录平台有中国哲学书电子化计划、汉典古籍、中国基本古籍库、中华经典古籍库等。各平台在数据收录方面各有特色，获取语料时应尽量选择数据准确、文献覆盖全面的数据库作为语料来源，同时也要注意将多个数据库语料相互对照、补充，以保证原始语料的准确性与完整性。

获取原始语料后需要对其进行数字化处理，首先以人工录入方式实现词表语料数字化，然后利用Python、R等程序在原始语料中完成词表候选语境的自动生成，再通过人工方式对候选语境进行修正，最终得到规范的语境数据。经过上述数字化处理后，将获得的词表数据与语境数据共同作为数字化文本分析对象。

2.2 分词与实体关系抽取

分词与实体关系抽取是分析流程的关键环节，其准确性直接影响分析结果的有效性。该部分主要包括文本自动分词与词性标注、实体标注与抽取、实体关系抽取、实体规范4个环节。

2.2.1 文本自动分词与词性标注

自动分词技术已经成为中文信息处理中最为基本和重要的研究内容，现有的自动分词技术主要包括基于规则的方法和基于统计的方法[10]。针对古籍文本的自动分词方法也在逐步发展和应用，如以反向最大匹配算法为主的多策略分词算法[11]、基于条件随机场的自动分词模型[12－13]等。词性标注主要是将语料库中的词按词性分类，常用的算法有最大熵马尔可夫模型、条件随机场等序列模型以及循环神经网络等深度学习算法。选择适用的自动分词与词性标注算法是保证分析结果可靠的重要前提。

2.2.2 实体标注与抽取

农作物时间分布与演化特征分析的目标实体有两类：作物实体与时间实体。首先随机选取部分语料对其中的两类实体进行人工标注，添加实体标签，然后以实体标注过的语料作为训练语料对模型进行训练，利用最优模型对目标语料进行实体自动识别与抽取。

(1)实体标注。一是作物实体标注。结合词表语料与原文语料确定作物实体词表,将其对应到原文中并添加作物标签,如在原文作物实体词的位置添加“<C>粟</C>”标签,C(crops,作物),作为后续自动抽取时的作物实体识别触发标记。二是时间实体标注。在词性标注的基础上,结合原文语义,人工对原文中标注有“/t”的时间类词语进行判读,补充修正漏标、错标的时间实体,并对时间类名词进行实体标注,标注形式为在时间表达式的位置添加“<T>绍兴五年</T>”标签,T(time,时间),作为后续自动抽取时的时间实体识别触发标记。

(2)实体自动识别与抽取。实体识别也称为命名实体识别,是自然语言处理中的一项重要任务,实体自动识别需要与实体标注相结合,才能达到较好的识别效果。在此用上文中已经添加过作物、时间实体标签的语料对机器学习模型进行训练,利用得到的最优模型进一步对目标语料进行实体自动识别与抽取。

2.2.3 实体关系抽取

完成上述步骤后,基于实体在原文中的语境特征和研究对象的自身属性制定合适的抽取规则,利用 Python、Java 等计算机语言调用内置函数自动化抽取不同实体的关联数据。抽取规则如下:首先以标签“<C>作物</C>”为触发词定位原文中的作物实体词,抽取该标签对应的作物实体;分别从作物实体的前置与后置位置定位与其字符距离最近的两个“<T>时间</T>”标签,依次抽取该标签对应的前后两个时间实体,抽取结果如表 1 所示:

表 1 实体关系抽取示例

作物	前置时间	后置时间
桑	‘#昔者;a2’	‘#正月始和;b1’
穀	‘#安帝永初三年;a2’	‘#桓帝永興元年;b1’
棗	‘#是時;a10’	‘#建安元年;b3’

表 1 中 a 代表前置距离,b 代表后置距离,字母后的数字表示该时间实体与对应作物实体的字符距离,数字越大,距离越大。最后通过人工方式结合原文语境对实体关系进行校对,从前置时间与后置时间中遴选出最为相关的时间作为与作物实体关联的时间实体,最终获得规范化的实体关联关系。

2.2.4 实体规范

就文中主要研究对象作物和时间而言,在古籍文本中农作物的类别存在着异名同指现象,在时间描述上亦有多种形式。为保证分析的准确性,研究进一步对自动抽取的实体数据进行规范。

(1)作物实体规范。对作物实体的规范主要是结合已有研究[14-15]与实际数据对农作物名称和类属进行统一、合并等,如将“菽”与“豆”合并为“豆”,将“粳”“糯”“秔”“稻”统一归为“稻”类。主要合并类别如表 2 所示:

表 2 主要作物实体规范

原文作物类	合并后	原文作物类	合并后	原文作物类	合并后
茶	茶	稻	稻	谷	谷
白牙		粳		糜	
先春		糯		秫	
探春		粇		鸡头	芡实
次春		秔		芡菱	
蒙顶		橘子	橘	黍	黍
水南		金橘		穄	黍
赵坡		温柑		粟	粟
荞	麦	豆	豆	桑	桑
麦		菽		枣	枣

(2)时间实体规范。根据原文文本中各时间表达式对时间描述的显隐性,本文将所有时间表达式划分为显式时间表达式与隐式时间表达式。显式时间表达式可以直接获得具体时间,隐式时间表达式则需要根据具体语境推测时间。本文结合语言学、历史学、文献学等方面的研究成果,对显式与隐式时间表达式的具体类型进行细分,如表 3 所示:

表 3 显式与隐式时间表达式类型

	类型	示例
显式时间表达式	年号+量词+年	永始二年
	王公/帝王+量词+年	孝文五年
	王公/帝王即位型	文帝即位
	天干地支型	太宗丙申年
隐式时间表达式	省略型(仅有年份,省略年号、王公/帝王等)	(紹興)四年
	指代型(用“是”“次”等代词指代时间)	是岁;次年
	方向型(用“前”“后”等词表示时间方向)	二年后
	区间型(在时间轴上表示为一个区间)	明朝初年至成化年间
	模糊型(无明确起止时间,在时间轴上无法定位)	十年间

然后,对时间实体进行规范,主要包括如下步骤:

第一步:对不同类型的显式与隐式时间表达式进行规范,规则如表 4 所示:

表 4　时间表达式规范规则

类别	规则	
显式时间表达式	年号＋量词＋年	对照年号年表获得朝代与公元纪年
	王公/帝王＋量词＋年	对照帝王年表获得朝代、年号与公元纪年
	王公/帝王即位型	对照帝王年表获得朝代、年号与公元纪年(即位元年)
	天干地支型	定位帝王或年号后将天干地支换算为具体年份，获得朝代、年号与公元纪年
隐式时间表达式	省略型	回溯原文语境，补全为显式时间表达式
	指代型	回溯原文语境，替换为显式时间表达式
	方位型	回溯原文语境，根据时间跨度换算为显式时间表达式
	区间型	回溯原文语境，转换为起始时间的显式时间表达式
	模糊型	由于无法定位具体时点，一般不做分析

第二步:对朝代、年号名称进行规范，如辽朝大康与太康均为辽道宗耶律洪基的同期年号的不同版本叫法，在此合并为辽－大康。同时，对朝代进行合并，如将“西汉”“新朝”及“东汉”合并为“汉朝”，将“北宋”“南宋”合并为“宋朝”，将“西晋”“东晋”合并为“晋朝”。

2.3　时间分布特征分析

为了揭示农作物在不同历史时期内的受关注程度与发展状况，首先对作物频次进行统计分析，然后构建作物－朝代、作物－年号、作物－公元纪年等不同时间维度上的关联数据矩阵，并利用 Ucinet、Gephi、Citespace 等可视化工具绘制知识图谱，结合节点属性及网络特征分析农作物的时间分布特征。

2.4　演化特征分析

为了更清晰地刻画农作物在时间轴上的发展演化趋势，需对农作物频次的动态变化情况进行时间序列分析与可视化。另外，农作物间存在着相互促进或相互排斥的影响[16－17]，从历史角度分析不同农作物是否具有相似的发展趋势，对于探索农作物之间可能的相互关系、全面把握我国农作物种植结构发展变化情况具有积极意义。这一目标主要通过分析农作物两两之间频次变化的相关性实现，常见的相关性分析系数有 pearson 系数、kendall 系数、spearman 系数等。

3　基于《食货志》文本的农作物时间分布及演化特征分析

3.1　数据准备

《食货志》是我国古代纪传体史书中专门叙述各代财政经济制度、农业生产、手工业发展实况等的志书，在记录历代经济发展现状与政策调整的同时，也涵盖了特定经济背景下农作物发展的重要信息和统计数据，是研究古代农业经济发展的重要知识来源。因此，本文以《食货志》文本作为分析对象。

首先选取《汉学引得丛刊》中的《食货志十五种综合引得》[18－19](以下简称《引得》)为词表语料，对其进行数字化录入与校对。同时从“中国哲学书电子化计划平台”与“汉典古籍”两个平台获取我国古代 15 本史书中《食货志》部分的原文文本(此处 15 本史书包括《史记》《汉书》《晋书》《魏书》《隋书》《旧唐书》《新唐书》《旧五代史》《宋史》《辽史》《金史》《元史》《新元史》《明史》《清史稿》，除《史记》中以《平准书》记录食货信息外，其余史书均设有《食货志》)，相互对照并结合后获得研究所需原文语料。然后利用 Python 程序对《引得》词表在原文语料中的语境进行自动抽取与人工校对。最终得到数字化的《引得》词表字(词)数 13 041个，原文语境数 191 946 条。

3.2　模型选取与效果评价

3.2.1　模型选取

2018 年谷歌提出一种基于自注意力机制(self-attention)建模的深度学习模型 BERT(Bidirectional Encoder Representations from Transformers)[20]，该模型舍弃了传统神经语言模型的循环神经网络结构，采用双向 Transformer 网络结构极大地提升了模型的特征提取能力。BERT 模型可以同时提取上下文信息，使得词语的表示具有更为准确和丰富的语义。应用在具体任务中，比如对某特定领域语料进行分词、词性标注[21]、实体识别[22]等任务时，该模型都取得了较好的实验效果，在不同类型典籍语料中的适用性得到了肯定。因此本文选择 BERT 模型对《食货志》原始语料进行分词词性一体化标注以及相关实体的识别和抽取。

3.2.2　效果评价

10 折交叉验证是机器学习中较为常用的验证模型有效性的方法，评价指标包括准确率 P(precision)、召回率 R(recall)和调和平均值 F(F-Measure)[22－25]。词性标注与实体识别的模型效果如表 5、表 6 所示:

表 5 BERT 模型在目标语料上的词性标注效果

（单位/%）

10 折编号	P	R	F
1	90.01	90.2	90.1
2	90.26	90.58	90.42
3	90.03	90.37	90.2
4	90.25	90.58	90.41
5	89.9	90.26	90.08
6	89.9	90.44	90.17
7	90.09	90.43	90.26
8	90.17	90.57	90.37
9	89.92	90.37	90.14
10	89.86	90.29	90.08
Average	90.04	90.41	90.22

从表 5 中可以看出在词性标注实验中，10 次交叉实验中 2 号语料效果最好，P、R、F 值略高于其他组实验结果，总体上 BERT 模型的词性标注评价结果，准确率达到了 90.04%，召回率达到了 90.41%，F 值达到了 90.22%，证明了 BERT 模型用于《食货志》语料上词性标注实验的有效性。

表 6 基于 BERT 模型的实体自动识别效果

（单位/%）

10 折编号	P	R	F
1	89.6	92.83	91.19
2	88.61	92.74	90.63
3	87.69	92.69	90.12
4	88.75	93.68	91.15
5	89.11	92.7	90.87
6	88.36	91.53	89.91
7	89.46	93.58	91.47
8	88.55	92.03	90.26
9	89.68	93.6	91.6
10	89.94	92.83	91.36
Average	88.98	92.82	90.86

从表 6 中可以看出在 10 次交叉验证结果中，整体 F 值达到了 90.86%，准确识别了绝大多数的目标实体类，模型与目标语料的契合性较为理想。整体上通过 BERT 模型实现《食货志》语料的分词词性一体化标注以及实体识别任务，10 折交叉验证结果证明了模型在目标语料任务上的可行性。

采用前文所述实体标注与实体关系抽取系列步骤，从 15 本《食货志》文本中抽取得到 2 366 条作物 - 朝代 - 年号 - 公元纪年关联数据，这也是后续进行数据统计和作物时间分布与演化分析的数据基础。

3.3 农作物时间分布特征分析

3.3.1 农作物频次统计分析

对规范化的作物实体进行统计，按照频次高低对作物实体词进行排序。考虑到不同史书《食货志》记录的详尽程度不同，为了消除不同时期文字记录水平带来的影响，提高分析的准确性与客观性，本文以相对频次衡量农作物发展热度。相对频次计算如公式（1）所示：

$$F_{rel}(C_1) = \alpha \times \sum_{t=1}^{n} \frac{f_t(C_1)}{W_t} \qquad 公式(1)$$

其中 $f_t(C_1)$ 为农作物 C_1 在文本编号为“t”的《食货志》语料中的频次，n 为样本总量，取值为 15，t 为按照史书编纂时间先后排列的原文文本序号，取值范围为 1 - n，W_t 为文本“t”的总字数，$F_{rel}(C_1)$ 为作物 C_1 的相对频次。即农作物的相对频次为该作物在每部《食货志》文本中的实际频次与对应文本总字数比值的和。由于农作物频次与文本字数之间的数量差异使得计算结果偏小不易于解读，本文将所有结果统一乘以系数 α，根据数据特征在此取值 10^5。经过计算得到农作物的相对频次（以下简称为频次）排序，Top10 统计结果如表 7 所示：

表 7 相对频次 Top10 作物

作物	绝对频次	相对值	相对频次	作物	绝对频次	相对值	相对频次
粟	480	0.015 712	1 571	稻	105	0.001 610	161
谷	324	0.013 310	1 331	豆	77	0.001 124	112
茶	990	0.010 205	1 020	黍	15	0.000 868	87
桑	146	0.005 012	501	枣	27	0.000 852	85
麦	135	0.002 379	238	棉	13	0.000 176	18

从表 7 中可以看出，“粟”的频次最高，要明显高于其他农作物，“粟”类作物不仅是古代中国尤其是北方地区主要的粮食作物，还广泛渗透于中国传统文化当中[26]。“谷”与“粟”的生长习性基本相同，也是北方地区的重要粮食作物之一，因此在《食货志》文本中的频次仅次于“粟”。排名紧随其后的是“茶”和“桑”，这两种作物是我国古代重要的经济作物，对我国古代商品经济的发展发挥着重要作用。“麦”和“稻”的频次也

较高,两者是具有比较典型的地域性特征的农作物,我国古代尤其隋唐以后,“北麦南稻”的局面逐渐形成[27]。“豆”“枣”“黍”“棉”等农作物的相对频次明显低于主要的粮食作物和经济作物,但也是古代农作物经济体系中不可缺少的组成部分,豆类曾在一定时期和区域内作为主要粮食作物而被种植[28]。本文将重点围绕上述几类农作物展开分析。

3.3.2 作物－朝代分布特征分析

本文选择 Ucinet 为可视化分析工具对作物－朝代关联数据进行直观呈现,结果如图 2 所示:

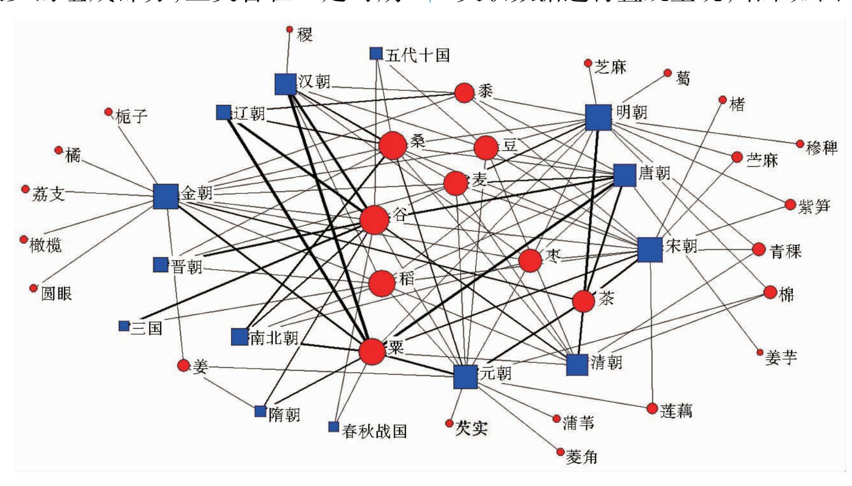


图 2 作物－朝代关联

图 2 中圆形节点表示农作物,方形节点表示朝代,节点与节点之间的连线表示农作物与朝代的关联关系,连线的粗细表示关联频次的大小。

从农作物角度,“粟”“茶”“谷”“麦”等与多个朝代都具有较高关联频次,说明这些农作物在古代经济史发展中发挥了关键作用。“粟”是我国最早的农作物之一,从春秋战国时期就已经成为经济发展的重要组成要素,还是政府俸禄制度的一部分[29]。“粟”在汉朝、唐朝、辽朝、金朝、元朝受关注程度均较高,其中与汉朝、辽朝的关联关系最为明显。汉朝晁错提出“重农贵粟”政策,促进了粟作农业的发展[30];辽立国之后便主张“专事于农”,促进了以旱地粮食作物为主的农业发展,“粟”成为种植范围最广的农作物[31－32]。“茶”在明朝、宋朝、唐朝、清朝等朝代发展较好。从图 2 中可以看出,“茶”与明朝的关联关系最高,这也受益于明朝改革贡茶制度、减轻茶税,促进了散叶茶的快速发展[33]。唐中期以后,中国“茶道”大行,宋承唐代饮茶之风,茶类种植日益普及[34]。“谷”与辽朝、汉朝、三国、晋朝、唐朝等朝代的关联频次较高。其中“谷”与辽朝的关联关系最为明显,一方面是因为辽朝统治者重视农耕,另一方面是“谷”的耐旱耐寒属性更适合逐渐转冷的辽朝统治区域[35]。“谷”与汉朝的关联频次也较高,这主要是汉初时期,重农抑商政策的大力推行,提高了大众的生产积极性。《氾胜之书》《四民月令》的编纂,也说明汉朝农业种植更加科学有序,“谷”类作物的种植业得到了进一步发展[36－37]。“麦”是起源于西亚,后传入中国的粮食作物[38]。从图 1 中可以看出“麦”与唐朝的关联频次最高,唐时麦类作物逐步上升到主流地位,成为北方最主要的粮食作物之一[39－40]。

从朝代角度,明朝、金朝、宋朝、元朝、唐朝、汉朝、清朝等朝代的关联农作物较多且关联频次更高。这些朝代在历史上存在的时间都较长,并且多数所处的时期是整个封建社会的中后期,可以充分借鉴前人农业发展经验,利农政策相对完善,为农业经济发展提供了良好的环境。为了进一步了解不同执政者统治时间段内的作物分布情况,本文将时间实体从朝代层面具体到年号层面进行分析。

3.3.3 作物－年号分布特征分析

提取作物－年号关系矩阵,利用 Ucinet 软件绘制作物－年号关联网络图谱,结果见图 3。

图 3 中圆形节点为农作物,方形节点为年号,节点间的连线表示作物与年号的关联关系,连线的粗细表示关联频次大小。为了使网络图谱更具可读性,调整可视化图的显示阈值为“ >7”,即仅展示关联频次在 7 次以上的关系。

从图 3 中可以看出,隋朝与农作物“粟”“谷”“姜”的关联频次较高,隋－开皇年间隋文帝杨坚推行系列利农政策,使得隋朝有了“计天下之储积,得供五六十年”的繁盛局面[41－42]。唐朝与“粟”“茶”“谷”“麦”“桑”“稻”等多种农作物都有较高的关联频次,特别是

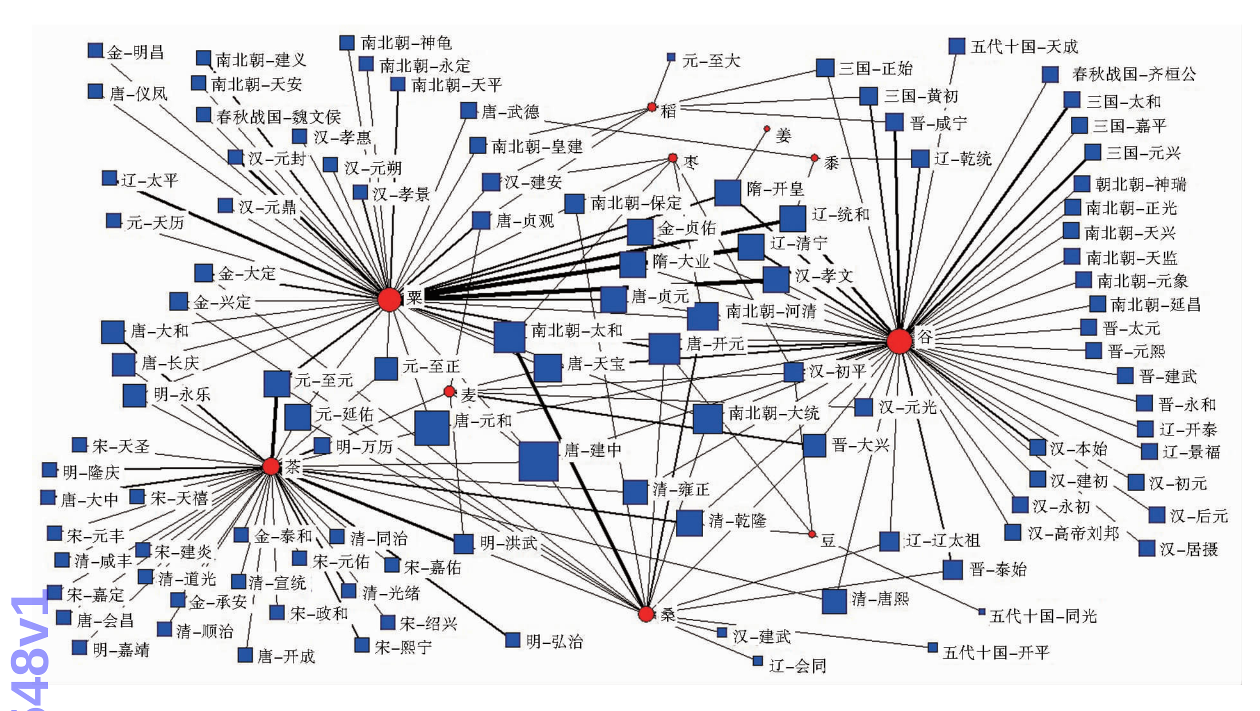


图 3　作物－年号关联

唐－开元年间唐玄宗李隆基采取了整顿吏治、兴修水利、改革户籍等多项举措,为农业经济的发展提供了有力保证,这才有了开元盛世的繁华局面[43－44]。辽朝与农作物“粟”“谷”“桑”“黍”等的关联频次较高,其中又以辽－统合、辽－清宁、辽－太平时期最为突出。辽时采取的“因俗而治”政策对辽的农业发展起到了重要的促进作用[45]。元朝与“茶”“粟”“桑”“稻”“麦”“谷”等农作物都有较高的关联频次,特别是元－至元时期元世祖忽必烈重视农业发展,从中央到地方建立专门的农司管理农业生产,颁布《农桑辑要》使农业种植更加科学合理有章法[46]。清朝与农作物“茶”“谷”“豆”“粟”等的关联频次较高,其中清－康熙、清－乾隆、清－雍正时期与农作物的关联关系最为明显。3 位皇帝采取了鼓励垦荒、放宽起科、更民田、摊丁入亩、兴修水利等措施,有效刺激了农业生产[47－49]。特别是乾隆皇帝最爱饮茶,茶叶的种植也在其统治期间得到了大范围的推广[50]。

3.4　农作物演化特征分析

3.4.1　演化趋势分析

本文进一步选取频次最高的前 8 种农作物,对其沿时间轴动态变化情况进行可视化分析,结果如图 4 所示:

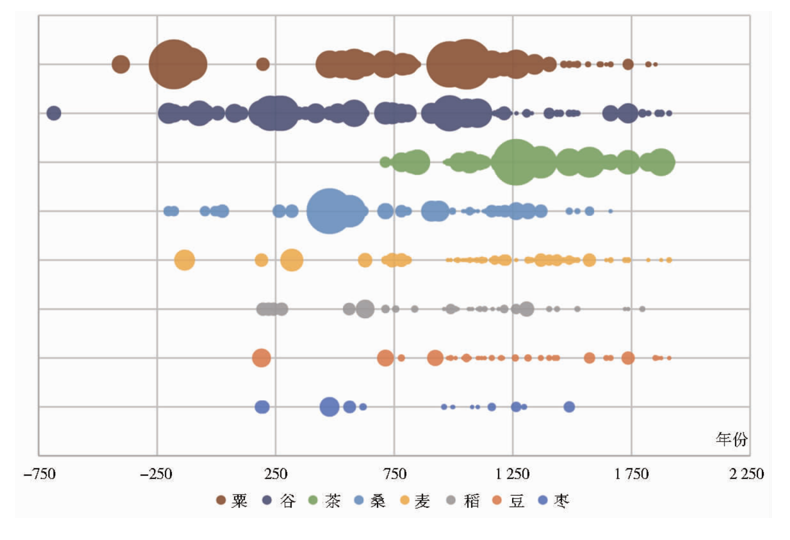


图 4　作物演化气泡图

图4中年份为横坐标，“粟”“谷”“茶”“桑”“麦”“稻”“豆”“枣”8种农作物自上而下沿纵坐标轴平均分布，相对频次高低以圆点面积大小表示。从整体来看，8种主要农作物在经济发展中开始受到重视的时间各不相同，后续发展态势及延续性也呈现出不同特征。“粟”和“谷”是较早被人们关注的两种农作物，春秋战国与秦汉时期的饮食结构以谷物为主，“粟”也日益占据主导地位，在汉代还成为口粮的代称[51]。而后这两种农作物迎来了较长时间的稳步发展，一直到宋金时期都具有良好的发展态势。但到了明清时期，这两种农作物比重及地位有所下降，这与“麦”“豆”等农作物比重提高以及域外粮食作物种植范围扩大具有一定的关联性[52]。总体而言，“粟”和“谷”两种农作物在其发展过程中具有较好的延续性，其发展脉络基本涵盖了我国历史发展的各个阶段，在我国古代经济社会发展中扮演着不可或缺的重要角色。“茶”的流行开始于唐朝，这一时期种茶开始从自然经济下的原始生产发展到商品经济下的社会生产[53]。根据陆羽的《茶经》记载，唐朝的制茶、煮茶、饮茶工具和技术已经十分成熟[54]。自此茶业得到快速发展，“茶”的发展主要经历了唐代初兴、宋代发展、明清鼎盛三大历史阶段[55]，其蓬勃发展态势对近现代茶业与茶文化的进步与发展具有积极影响。“桑”大致兴起于春秋战国时期[56]，在其发展前期波动性较大，后又经历了较长时间的平稳发展阶段。但到了明清时期，蚕桑业受到赋税制度改革和棉花崛起的影响，至清末趋于衰落[57]。“麦”“稻”“豆”和“枣”则在秦汉时期成为人们的主要粮食与水果、干果[58]，在经历了一段时间的起伏波动后自唐宋开始逐步趋于稳定。整体来看，“桑”“麦”“稻”“豆”和“枣”的发展虽不及“粟”“谷”强势，但同样具有良好的发展态势和延续性，因而也是古代农业经济社会发展的重要推动元素。

3.4.2 演化相关性分析

本文选取频次大于4的农作物，对其两两之间的频次变化情况做 pearson、kendall、spearman 相关性分析。将分析结果与大量史料对比后发现，pearson 相关性分析结果更符合历史发展实际，对农作物相互关系的表征性更强，因此最终选择 pearson 系数作为农作物相关性分析指标。对分析结果进行数据可视化，如图5所示。图中正相关用蓝色表示，负相关用红色表示，颜色越深表示相关性越强。

	粟	谷	茶	桑	麦	稻	豆	黍	枣	姜	棉	楮	橘	稷	姜芋	莲藕	青稞
粟	1.00	0.30	0.03	0.15	-0.01	0.03	0.00	0.23	0.13	0.16	0.03	-0.07	-0.02	-0.03	0.01	0.13	-0.04
谷	0.30	1.00	-0.13	0.07	0.04	0.07	0.13	0.32	0.05	0.14	-0.09	-0.08	-0.04	0.03	-0.04	-0.05	0.00
茶	0.03	-0.13	1.00	0.05	0.15	0.05	0.13	-0.06	0.11	0.05	0.52	-0.01	0.03	-0.03	0.12	0.62	0.04
桑	0.15	0.07	0.05	1.00	0.02	-0.01	0.02	-0.01	0.76	-0.01	0.06	-0.04	-0.01	-0.02	-0.02	0.10	-0.03
麦	-0.01	0.04	0.15	0.02	1.00	0.09	0.21	-0.04	0.02	-0.03	0.13	0.01	-0.03	-0.03	-0.03	-0.01	0.12
稻	0.03	0.07	0.05	-0.01	0.09	1.00	0.03	0.00	0.07	-0.03	0.10	0.00	-0.03	-0.03	-0.03	0.14	0.02
豆	0.00	0.13	0.13	0.02	0.21	0.03	1.00	-0.03	0.16	-0.03	0.00	0.00	-0.02	-0.02	-0.02	-0.02	0.15
黍	0.23	0.32	-0.06	-0.01	-0.04	0.00	-0.03	1.00	0.02	-0.02	-0.03	-0.03	-0.01	0.13	-0.01	-0.01	-0.02
枣	0.13	0.05	0.11	0.76	0.02	0.07	0.16	0.02	1.00	-0.01	0.14	-0.04	-0.02	-0.02	-0.02	0.19	-0.03
姜	0.16	0.14	0.05	-0.01	-0.03	-0.03	-0.03	-0.02	-0.01	1.00	0.11	-0.02	0.38	-0.01	-0.01	0.08	-0.01
棉	0.03	-0.09	0.52	0.06	0.13	0.10	0.00	-0.03	0.14	0.11	1.00	-0.04	-0.01	-0.01	-0.01	0.31	-0.02
楮	-0.07	-0.08	-0.01	-0.04	0.01	0.00	0.00	-0.03	-0.04	-0.02	-0.04	1.00	-0.01	-0.01	-0.01	-0.02	-0.02
橘	-0.02	-0.04	0.03	-0.01	-0.03	-0.03	-0.02	-0.01	-0.02	0.38	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01
稷	-0.03	0.03	-0.03	-0.02	-0.03	-0.03	-0.02	0.13	-0.02	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01
姜芋	0.01	-0.04	0.12	-0.02	-0.03	-0.03	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01
莲藕	0.13	-0.05	0.62	0.10	-0.01	0.14	-0.02	-0.01	0.19	0.08	0.31	-0.02	-0.01	-0.01	-0.01	1.00	-0.01
青稞	-0.04	0.00	0.04	-0.03	0.12	0.02	0.15	-0.02	-0.03	-0.01	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01	1.00

图5　作物演化相关性

从图5中可以看出，农作物发展态势之间存在正、负相关两种关系，但相关关系较为显著的几组数据均为正相关，所有负相关关系都不显著，这说明这些农作物组合之间更可能存在相互促进关系。相关性最强的一组农作物为“桑－枣”，古代诗歌中“江上数株桑枣树”“前种桑麻后梨枣”等诗句中较好地佐证了两者的相关性。这与统治者的治国思路有重要关系，特别是在元明时期，忽必烈与朱元璋都采取了鼓励百姓同时种植桑枣的重要举措[59－60]。“莲藕－茶”“棉－茶”的相关性也较强，这体现出“莲藕”与“茶”“棉”与“茶”的发展历程较为类似。具体而言，“莲藕”在唐时的栽培出现了一个较为活跃和集中的时期，后在宋元时期不断发展，到明清时期其种植栽培、生产、加工、利用等各项活动臻于成熟[61]；“棉”在唐宋时期在我国边远地

区的种植态势良好,元朝初年在长江以南得到较大发展,明清时代逐步形成华北棉区和华中棉区,植棉纺织生产也在全国各地广泛发展[62]。除这 3 组数据之外,其他农作物组合之间的相关性都偏低,相关关系均不十分显著。

4 结语

本文提出了一套完整的基于典籍文本的农作物时间分布及演化特征分析方法流程,根据典籍文献的语言组织特征,提出显隐性时间表达式的类别划分和具体的规范方式,可以为典籍文本挖掘中时间实体识别与清洗提供参考;利用深度学习模型 BERT 实现对原始语料的分词词性一体化标注,并以添加人工标签的语料对实体标注与实体关系自动抽取模型进行训练,进而实现对《食货志》文本相关实体的自动识别和抽取,可以为典籍文本的知识抽取任务提供借鉴;融合时间序列分析、知识图谱分析等方法和技术,实现了对多类型农作物时间分布及演化特征的定量化与可视化分析,可以为农史类科学研究提供新思路。研究将该方法应用于《食货志》典籍文本,分析结果得到了历史学、经济学、文献学等多学科相关研究资料的佐证,对方法的可行性与有效性进行了验证。本文研究是在数字人文视角下进行古文本情报挖掘和利用的一次实践探索,该方法流程可以拓展到《氾胜之书》《四民月令》《农桑辑要》《农书》等更多记录我国古代农作物发展情况的典籍文献研究中。

未来研究还可以做如下改进:①本文采用的方法流程自动化水平有待提高,后续将训练更多自动化模型,进一步提升分析效率;②分析可能具有一定的样本依赖性,后续将逐步扩大研究样本,以加强对方法有效性的验证;③典籍文本中对农作物的记载可能包括多种事件类型,如农作物种植、农作物加工、农作物用具制造等,未来研究可考虑进一步细分事件类型,以更加全面地分析古代农作物的多维特性。还需要注意的是,这种自动化分析技术虽然提高了分析效率,却不能完全取代传统的人工解读方法,分析时需将两者有效结合,以兼顾分析的准确性与高效性。

参考文献:

[1] 新华网. 习近平出席中央农村工作会议并发表重要讲话[EB/OL].[2021-05-15]. http://www.cppcc.gov.cn/zxww/2020/12/30/ARTI1609288702470104.shtml,2020-12-30.

[2] 陈明远,金岷彬. 历史考古的新观点(之十) 甲骨文中的谷类及东西方谷物加工技术的比较研究[J]. 社会科学论坛,2014(10):16-35.

[3] 李成. 黄河流域史前至两汉小麦种植与推广研究[D]. 西安:西北大学,2014.

[4] 刘兴林. 先秦两汉农作物分布组合的考古学研究[J]. 考古学报, 2016 (4): 465-494.

[5] 简思敏,刘锡涛. 福建明清时期农作物的地理分布[J]. 福建地理, 2005(4): 50-54.

[6] 李静. 清至民国北川地区主要农作物的种植及其分布[J]. 古今农业, 2009(2): 77-83.

[7] 朱睿,杨飞,周波,等. 中国苎麻的起源、分布与栽培利用史[J]. 中国农学通报, 2014, 30(12): 258-266.

[8] 周跃中. 试谈中国古代农作物种类及其历史演变[J]. 吉林农业, 2010(8): 1-3.

[9] 彭景元. 闽南古代农业述略[J]. 古今农业, 2005(2): 11-25.

[10] 黄水清,王东波,何琳. 以《汉学引得丛刊》为领域词表的先秦典籍自动分词探讨[J]. 图书情报工作,2015,59(11):127-133.

[11] 邱冰,皇甫娟. 基于中文信息处理的古代汉语分词研究[J]. 微计算机信息,2008(24):100-102.

[12] 石民,李斌,陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究[J]. 中文信息学报,2010,24(2):39-45.

[13] 王姗姗,王东波,黄水清,等. 多维领域知识下的《诗经》自动分词研究[J]. 情报学报,2018,37(2):183-193.

[14] 谢月涵.《说文解字》"食"部字与饮食文化探究[J]. 绵阳师范学院学报,2021,40(1):79-85,91.

[15] 梁欢. 从《说文解字》禾部字看中国古代的农业文化[J]. 黑河学院学报,2020,11(4):165-169.

[16] 张如义,王仕林,胡红玲,等. 3 种作物(莴笋、茄子、小白菜)对香樟凋落叶化感作用的生理响应[J]. 热带亚热带植物学报,2021,29(1):41-49.

[17] RICE E L,王天伦. 农作物的植物型间生物化学相互作用[J]. 耕作与栽培,1989(2):50-53,55.

[18] 马学良,孙蕊. 从"整理国故"看哈佛燕京学社汉学引得丛刊的价值[J]. 图书情报工作,2010,54(7):111-114.

[19] 马学良,李伟. 哈佛燕京学社汉学引得丛刊的文献学价值与思想[J]. 河北大学学报(哲学社会科学版),2010,35(2):94-98.

[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[21] 张琪,江川,纪有书,等. 面向多领域先秦典籍的分词词性一体化自动标注模型构建[J]. 数据分析与知识发现,2021,5(3):2-11.

[22] 杜悦,王东波,江川,等. 数字人文下的典籍深度学习实体自动识别模型构建及应用研究[J/OL]. 图书情报工作:1-9[2021

－04－09]. https://doi.org/10.13266/j.issn.0252-3116.2021.03.013.
[23] 闫壮壮,闫学慧,石嘉,等. 基于深度学习的大豆豆荚类别识别研究[J]. 作物学报,2020,46(11):1771－1779.
[24] 陈永超. 基于机器学习的心音分类算法研究[D]. 济南:山东大学,2020.
[25] 宋亚斌,邢元军,江腾宇,等. 基于距离相关系数和 KNN 回归模型的森林蓄积量估测研究[J]. 中南林业科技大学学报,2020,40(4):22－27,33.
[26] 高强. 粟与粟文化[J]. 华夏文化,2003(4):15－17.
[27] 周跃中. 试谈中国古代农作物种类及其历史演变[J]. 吉林农业,2010(8):1－3.
[28] 杨坚. 古代大豆作为主食利用的研究[J]. 古今农业,2000(2):16－22.
[29] 阎步克. 从稍食到月俸——战国秦汉禄秩等级制新探[J]. 学术界,2000(2):61－82.
[30] 汤标中. 文景之治与积贮贵粟[J]. 中国粮食经济,1999(6):43－46.
[31] 李钧. 辽国农业的发展[J]. 西南民族学院学报(哲学社会科学版),1990(3):80－84.
[32] 韩茂莉. 辽代农作物地理分布与种植制度[J]. 中国农史,1998(4):22－29.
[33] 沈志荣. 明代"茶商为神"探究[J]. 杭州(周刊),2018(16):56－57.
[34] 李昕升,王思明. 评《中国古代粟作史》——兼及作物史研究展望[J]. 农业考古,2015(6):341－343.
[35] 田晓雷. 辽代饮食结构新探[J]. 阴山学刊,2015,28(5):74－80.
[36] 陈冬仿. 汉代农业生产的生态意蕴[J]. 中州学刊,2019(11):121－124.
[37] 郭建新. 汉代农业科技政策与管理探析[J]. 商丘师范学院学报,2019,35(5):75－79.
[38] 赵志军. 小麦传入中国的研究——植物考古资料[J]. 南方文物,2015(3):44－52.
[39] 华信辉.《全唐诗》中的唐代麦类作物及其影响[J]. 三明学院学报,2017,34(1):82－85.
[40] 包艳杰,李群. 唐宋时期华北冬小麦主粮地位的确立[J]. 中国农史,2015,34(1):49－58.
[41] 梁方仲. 论隋代经济高涨的原因[J]. 历史教学,1956(12):10－16.
[42] 严黎明. 浅论隋代国富之原因[J]. 西北成人教育学院学报,2019(2):84－89.
[43] 陈秀平,欧阳庆芳. 隋唐时期农业立法及农业发展状况浅析[J]. 法制与社会,2009(14):365－366.
[44] 马旭. 开元盛世经济繁荣的原因分析[J]. 才智,2017(26):203.
[45] 于金华. 简论辽朝的"因俗而治"政策[J]. 自贡师专学报,1998(2):13－18.
[46] 海日. 论元世祖忽必烈的经济政策[J]. 前沿,2009(5):75－77.
[47] 曹巧. 论清朝前期环北部湾地区的农业垦殖[J]. 湛江师范学院学报,2011,32(4):111－114.
[48] 柏林. 雍正:康乾盛世的有力推行者[J]. 人才资源开发,2015(3):103－104.
[49] 肖婷. 试析农业发展对"康乾盛世"稳固所起的作用[J]. 农业考古,2012(1):74－78,106.
[50] 李幸哲,宋时磊. 乾隆八旬万寿庆典与清代宫廷茶文化——以朝鲜徐浩修《燕行纪》为中心[J]. 农业考古,2020(2):15－21.
[51] 陈文华. 春秋战国、秦汉时期的饮食文化[J]. 农业考古,2007(4):236－246,248－249.
[52] 李秋芳. 明清时期华北平原粮食种植结构变迁研究[M]. 北京:社会科学文献出版社, 2016.
[53] 吕维新. 唐代茶叶生产发展和演变[J]. 茶叶通讯,1989(4):53－54,57.
[54] 陆羽,钟强. 茶经[M]. 哈尔滨:黑龙江科学技术出版社, 2012.
[55] 胡长春. 明清时期中国茶文化的变革与发展[J]. 农业考古,2012(5):18－26.
[56] 吴琼. 秦汉蚕桑丝织技术和早期丝绸之路[J]. 科学技术哲学研究,2015,32(1):75－81.
[57] 马雪芹. 明清河南桑麻业的兴衰[J]. 中国农史,2000(3):53－56,72.
[58] 刘尊志. 秦汉三国时期食物的品种[J]. 大众考古,2017(2):94－95.
[59] 人民教育出版社,课程教材研究所,历史课程教材研究开发中心. 历史[M]. 北京:人民教育出版社, 2007.
[60] 张显清. 明太祖朱元璋社会理想、治国方略及治国实践论纲[J]. 明史研究,2007:6－44.
[61] 曹蓓蓓,丁晓蕾. 中国古代莲藕栽培起源概说[J]. 绿色科技,2015(12):129－131.
[62] 史学通,周谦. 元代的植棉与纺织及其历史地位[J]. 文史哲,1983(1):35－45.

作者贡献说明:
崔斌:数据处理与论文撰写;
王东波:论文撰写与修改指导;
黄水清:论文选题指导与审阅。

The Analysis of Time Distribution and Evolution Characteristics of Crops in Classics: Taking *Shihuozhi* as an Example

Cui Bin[1,2]　Wang Dongbo[1,2]　Huang Shuiqing[1,2]
[1] College of Information Management, Nanjing Agricultural University, Nanjing 210095
[2] Research Center for Humanities and Social Computing, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] There is a long history of crop cultivation in China. It is of great significance to analyze the time distribution and development evolution of ancient crops for optimizing the modern agricultural planting structure. [**Method/process**] This paper put forward a set of analytical process of crop time distribution and evolution characteristics, which included four parts: corpus acquisition and digitization, segmentation and entity relationship extraction, time distribution characteristics analysis and evolution characteristics analysis, and selected *Shihuozhi* from 15 historical books for empirical analysis. [**Result/conclusion**] Based on the analysis results of *Shihuozhi*, the feasibility and effectiveness of the method are verified by the relevant historical, economic, philological and other multidisciplinary research data, which can provide reference for the analysis of the time distribution and evolution characteristics of ancient crops based on classics. But in the future, we need to improve the level of automation, expand the research sample, refine the event type and other aspects to further optimize the method process.

Keywords: entity association　digital humanities　*Shihuozhi*　crops　visualization